III - Persistent homology

PSL Week - Topological Data Analysis

Abstract

We explain how to track the homology groups to a whole *family* of spaces simultaneously, and how to summarize the result by a simple object with stability properties: the *persistence diagram*.

In a typical TDA pipeline, we start from a point cloud in a metric space, build a nested family (a *filtration*) of simplicial complexes, compute homology at all scales, and visualize how connected components, loops and higher-dimensional holes appear and disappear as the scale changes.

Contents

1 Filtrations of spaces and complexes

1.1 Filtrations

The idea of a filtration is to look at a space X through a family of nested subsets, ordered by a parameter, often interpretable as a scale or time.

Definition 1.1 (Filtration of a space)**.** Let X be a topological space. A *filtration* of X indexed by a totally ordered set $(T, \leq)$ (typically $T = \mathbb{R}$ or $T = \mathbb{N}$) is a family of subspaces $(X_t)_{t\in T}$ such that:

(i) $X_s \subset X_t$ whenever $s \leq t$;

(ii) $\bigcup_{t\in T} X_t = X$.

Example 1.2 (Sublevel sets of a function)**.** Let $f : X \to \mathbb{R}$ be continuous. For each $t \in \mathbb{R}$, define the *sublevel set*

$$X_t := \{x \in X : f(x) \leq t\}.$$

Then $(X_t)_{t\in\mathbb{R}}$ is a filtration: if $s \leq t$ then $X_s \subset X_t$, and $\bigcup_t X_t = X$. For instance, one can take f to be the height function on a surface embedded in $\mathbb{R}^3$.

Example 1.3 (Offset filtration)**.** Let X be a compact subset of a metric space (M, d). Write

$$\mathrm{dist}(y, X) := \inf_{x \in X} d(x, y)$$

for the distance from y to set X. For $t \geq 0$ define the thickening

$$X_t := \{y \in M : \mathrm{dist}(y, X) \leq t\}.$$

Then $(X_t)_{t \geq 0}$ is a filtration. X_t is call the t-offset of X in M. See Figure 1 for X being a finite sample.

1.2 Filtrations of simplicial complexes

In computations we usually work with simplicial complexes rather than arbitrary spaces.

Definition 1.4 (Filtration of complexes)**.** A *filtration of simplicial complexes* is a family $(\mathcal{K}_t)_{t \in T}$, where each $\mathcal{K}_t$ is a simplicial complex and

$$s \leq t \implies \mathcal{K}_s \subset \mathcal{K}_t$$

(as subcomplexes, i.e. at the level of simplices).

Example 1.5 (Čech and Vietoris–Rips filtrations)**.** Let $\mathcal{P} \subset \mathbb{R}^d$ be a finite point cloud. For each $\alpha > 0$ we defined in Chapter 2:

- the Čech complex $\mathrm{Cech}(\mathcal{P}, \alpha)$,
- the Vietoris–Rips complex $\mathrm{Rips}(\mathcal{P}, \alpha)$.

As the scale α increases, these complexes are nested:

$$\alpha \leq \beta \implies \mathrm{Cech}(\mathcal{P}, \alpha) \subset \mathrm{Cech}(\mathcal{P}, \beta), \qquad \mathrm{Rips}(\mathcal{P}, \alpha) \subset \mathrm{Rips}(\mathcal{P}, \beta).$$

Thus $(\mathrm{Cech}(\mathcal{P}, \alpha))_{\alpha > 0}$ and $(\mathrm{Rips}(\mathcal{P}, \alpha))_{\alpha > 0}$ are filtrations. These are the main constructions used in TDA.

Example 1.6 (Finite filtrations)**.** In many practical situations we consider a *finite* increasing sequence of complexes

$$\mathcal{K}_0 \subset \mathcal{K}_1 \subset \cdots \subset \mathcal{K}_m,$$

for example obtained by sorting the simplices by some "time of appearance" (distance threshold, function value, etc.). We may take $T = \{0, 1, \ldots, m\}$ with the usual order.

2 Persistence homology structures

Given a filtration $(\mathcal{K}_t)_{t \in T}$ of simplicial complexes, we can form homology at each index t to get a family of vector spaces $H_k(\mathcal{K}_t)$ for each fixed dimension k. Because the complexes are nested, we also get linear maps between these spaces, induced by the inclusions. This structure is called a *persistence module*. It allows to keep track of how the sequence of homology groups evolves as a whole.

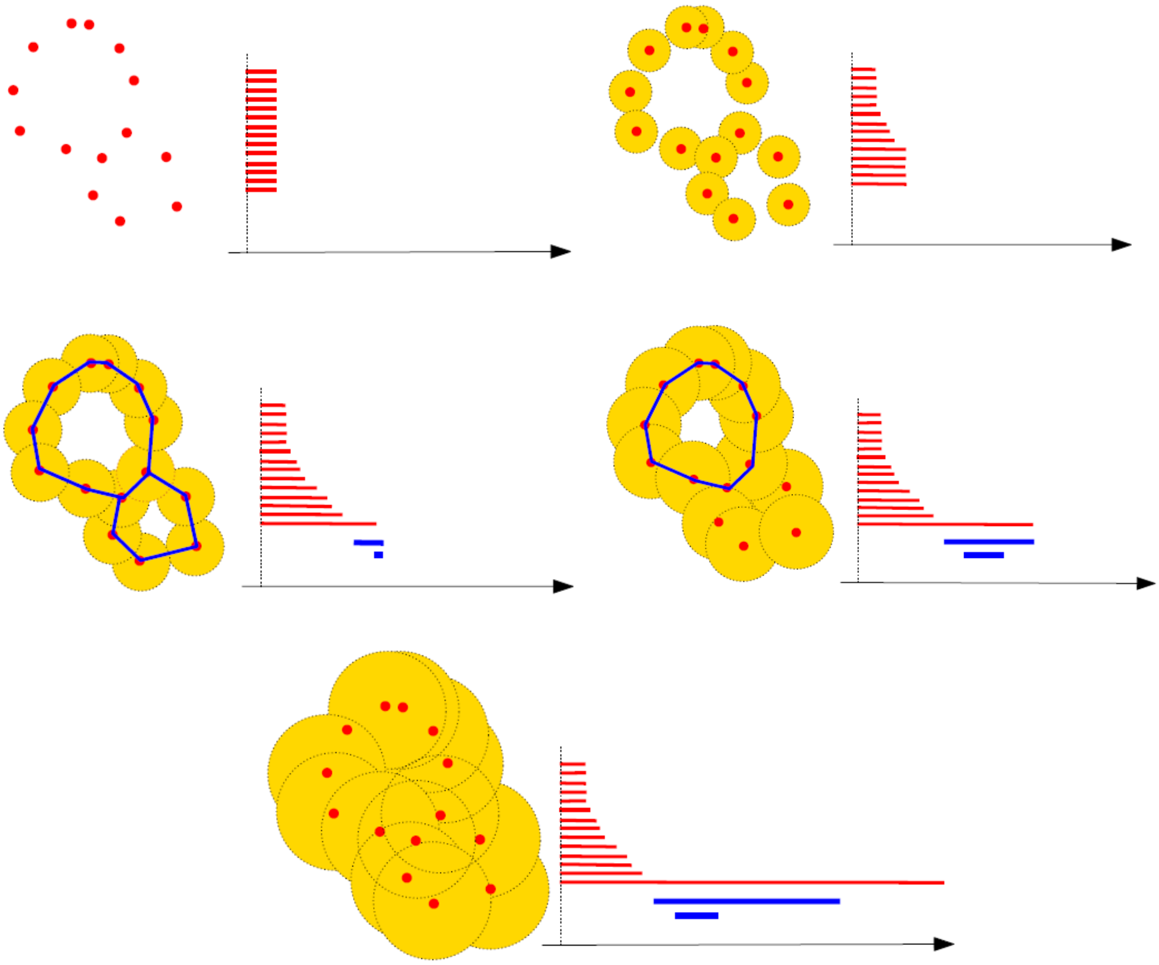

Figure 1: A point cloud on a circle and its Cech filtration at increasing scales α, with associated barcodes of dimensions 0 (red) and 1 (blue).

2.1 Modules

Throughout this section we work over a fixed field $\mathbb{F}$, often $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$ as in Chapter 2.

Definition 2.1 (Persistence module)**.** Let $(T, \leq)$ be a totally ordered set. A *(homological) persistence module* over T (with coefficients in $\mathbb{F}$) is:

- a family of $\mathbb{F}$-vector spaces $(V_t)_{t\in T}$,
- for all $s \leq t$ linear maps $\varphi_s^t : V_s \to V_t$

such that:

(i) $\varphi_t^t = \mathrm{id}_{V_t}$ for all $t \in T$;

(ii) for all $r \leq s \leq t$,

$$\varphi_s^t \circ \varphi_r^s = \varphi_r^t.$$

You should think of V_t as the homology at "time" t, and φ_s^t as telling how classes at time s evolve when we go forward to time t.

Example 2.2 (Homology of a filtration)**.** Let $(\mathcal{K}_t)_{t\in T}$ be a filtration of simplicial complexes. Fix a dimension $k \geq 0$.

For each t, let

$$V_t := H_k(\mathcal{K}_t; \mathbb{F}).$$

For $s \leq t$, the inclusion $\iota_s^t : \mathcal{K}_s \hookrightarrow \mathcal{K}_t$ induces a linear map on homology

$$(\iota_s^t)_* : H_k(\mathcal{K}_s) \to H_k(\mathcal{K}_t).$$

Set $\varphi_s^t := (\iota_s^t)_*$. Functoriality of homology implies that this family satisfies the axioms of a persistence module. This is the central example in TDA.

$$\{0\} \xrightarrow{\varphi_0^1 = (0)} \mathbb{F} \xrightarrow{\varphi_1^2 = \begin{pmatrix}1\\0\end{pmatrix}} \mathbb{F}^2 \xrightarrow{\varphi_2^3 = \begin{pmatrix}0 & 1\end{pmatrix}} \mathbb{F}$$

$$\{0\} \xrightarrow{\varphi_0^1 = (0)} \mathbb{F} \xrightarrow{\varphi_1^2 = \begin{pmatrix}1\\0\end{pmatrix}} \mathbb{F}^2 \xrightarrow{\varphi_2^3 = \begin{pmatrix}1 & 0\end{pmatrix}} \mathbb{F}$$

Figure 2: Two persistence modules. The one on the top has interval decomposition $\{[1, 2], [2, 3]\}$, while it is $\{[1, 3], [2, 2]\}$ for the bottom one.

2.2 Barcodes

Definition 2.3 (Interval module)**.** Let $I \subset T$ be an interval (for instance $I = [a, b)$, or $I = [a, \infty)$ if $T \subset \mathbb{R}$). The *interval module* $\mathbb{F}^I$ is the persistence module defined by:

- $V_t = \mathbb{F}$ if $t \in I$, and $V_t = 0$ otherwise;
- for $s \leq t$, φ_s^t is:

$$\varphi_s^t = \begin{cases} \mathrm{id}_{\mathbb{F}} & \text{if } s, t \in I, \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, the module is "on" (one-dimensional) exactly on the interval I, and zero elsewhere.

Theorem 2.4 (Structure theorem (informal)). *Under reasonable finiteness assumptions (which are satisfied for homology of finite filtrations), any persistence module V over $T \subset \mathbb{R}$ can be decomposed (non-uniquely as a module, but uniquely up to isomorphism) as a direct sum of interval modules:*

$$V \cong \bigoplus_j \mathbb{F}^{I_j}.$$

The multiset of intervals $\{I_j\}$ is called the barcode *of V.*

We will not prove this theorem; instead, we will use it as a guiding picture: each homology class is born at some parameter value (when it appears) and dies at a later value (when it becomes a boundary). Each such class corresponds to an interval $[b, d)$ in the barcode.

2.3 Diagrams

Barcodes are collections of intervals $[b, d)$ (birth and death times). For interpretation and visualization, it is often convenient to encode them as a discrete multiset of points (or measure) in the plane.

Remark 2.5 (Tameness). In applications, V_t is finite-dimensional for each t. Such persistence modules with this property are called *tame*. Under tameness *only*, no barcode decomposition can be obtained generically, however, the weaker notion of *persistence diagrams* (see below) is still well-defined.

Definition 2.6 (Persistence diagram as multisets). Let V be a persistence module decomposed into interval modules $\mathbb{F}^{[b_j,d_j)}$ and $\mathbb{F}^{[b_j,\infty)}$. The *kth persistence diagram* (for a fixed homological degree) associated to V is the multiset of points in the extended plane

$$\mathrm{Dgm}(V) := \{(b_j, d_j) \in \mathbb{R}^2 : \text{finite intervals}\} \cup \{(b_j, \infty) : \text{infinite intervals}\},$$

where each interval $[b_j, d_j)$ contributes one point of multiplicity 1, and similarly for $[b_j, \infty)$.

In practice, one usually plots only finite points (b_j, d_j) in the half-plane $\{(b, d) \mid b < d\}$, sometimes truncating very long intervals, and remembers separately the number of infinite bars. Equivalently, one can see persistence diagrams as purely discrete measures.

Definition 2.7 (Persistence diagram as measures). Let V be a persistence module decomposed into interval modules $\mathbb{F}^{[b_j,d_j)}$ and $\mathbb{F}^{[b_j,\infty)}$. The *kth persistence diagram* (for a fixed homological degree) associated to V is the measure on the extended plane

$$\mathrm{Dgm}(V) := \sum_j \delta_{(b_j,d_j)},$$

3 Bottleneck distance between diagrams

To compare the shapes of two datasets, or the effect of noise, we need a way to compare persistence diagrams. The standard metric is the *bottleneck distance*.

3.1 Definition

Let D_1 and D_2 be two persistence diagrams (thought of as multisets of points in the open half-plane $\{(b, d) \mid b < d\}$). Following the usual convention, we also allow matching points in one diagram to points on the diagonal $\{(x, x)\}$, representing intervals of length 0 (i.e. classes that are born and die instantly).

Definition 3.1 (Bottleneck distance)**.** The *bottleneck distance* between diagrams D_1 and D_2 is

$$d_B(D_1, D_2) := \inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_\infty,$$

where:

- the infimum is taken over all bijections $\gamma : D_1 \cup \Delta \to D_2 \cup \Delta$, where $\Delta = \{(x, x) : x \in \mathbb{R}\}$ is the diagonal with infinite multiplicity,
- $\| \cdot \|_\infty$ is the maximum norm: $\|(b_1, d_1) - (b_2, d_2)\|_\infty = \max(|b_1 - b_2|, |d_1 - d_2|)$.

Intuitively:

- We match points of D_1 to points of D_2, or if needed to the diagonal (interpreted as "noise" that can be killed).
- The cost of a matching is the largest shift in birth or death time needed to align corresponding points (measured in $\| \cdot \|_\infty$).
- We take the infimum over all matchings.

Remark 3.2. Matching a point (b, d) to the diagonal (x, x) costs at least $\frac{d-b}{2}$, since the closest diagonal point is $\left(\frac{b+d}{2}, \frac{b+d}{2}\right)$. Thus killing a long interval is expensive, while killing a very short interval is cheap.

4 Stability of persistence

A crucial property of persistence diagrams is that they are *stable*: small perturbations of the input produce only small changes in the diagrams, measured by the bottleneck distance.

4.1 Stability of sublevel set filtrations

Let X be a compact metric space, and let $f, g : X \to \mathbb{R}$ be two continuous functions. For each $t \in \mathbb{R}$ define the sublevel sets

$$X_t^f := \{x \in X : f(x) \le t\}, \qquad X_t^g := \{x \in X : g(x) \le t\}.$$

For a fixed homological degree k, these define two persistence modules:

$$t \mapsto H_k(X_t^f), \qquad t \mapsto H_k(X_t^g),$$

and two associated persistence diagrams D_k^f and D_k^g.

Theorem 4.1 (Stability of persistence diagrams for functions)**.** *Let X be a compact metric space and $f, g : X \to \mathbb{R}$ continuous. For each $k \ge 0$,*

$$d_B(D_k^f, D_k^g) \ \le\ \|f - g\|_\infty,$$

where $\|f - g\|_\infty = \sup_{x \in X} |f(x) - g(x)|$.

Idea. If $\|f - g\|_\infty \le \varepsilon$, then for every t:

$$X_t^f \subset X_{t+\varepsilon}^g \quad \text{and} \quad X_t^g \subset X_{t+\varepsilon}^f.$$

This gives, for each k, a family of linear maps between the persistence modules shifted by ε in both directions. One says that the corresponding modules are *ε-interleaved.* The classification of persistence modules and the definition of bottleneck distance imply that such an interleaving forces the diagrams to be at bottleneck distance at most ε. □

Remark 4.2. The precise notion of interleaving is algebraic, but the geometric intuition is simple: if f and g differ by at most ε, then the sublevel sets at level t for one function are contained in the sublevel sets at level $t + \varepsilon$ for the other. Thus the birth and death times of homology classes can only shift by at most ε, which is exactly what the bottleneck distance measures.

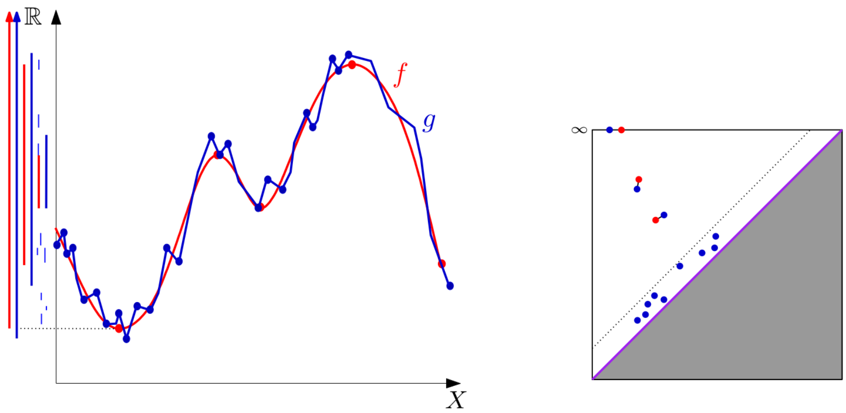

Figure 3: Zeroth order persistent homology of the sublevel sets of two closeby functions $f, g : [0,1] \to \mathbb{R}$.

4.2 Stability for point clouds

There are also stability results comparing the persistence diagrams of Rips or Čech filtrations built on finite subsets of a metric space, with respect to the Hausdorff distance between point clouds. We state one very informal version.

Theorem 4.3 (Informal stability for Rips filtrations)**.** *Let P, Q be two finite subsets of a metric space (M, d), and let $d_H(P, Q)$ be their Hausdorff distance. Then, for each k, the persistence diagrams of the Rips filtrations* $\mathrm{Rips}(P, \alpha)$ *and* $\mathrm{Rips}(Q, \alpha)$ *(built with the same metric d) satisfy*

$$d_B\big(D_k^{\mathrm{Rips}}(P), D_k^{\mathrm{Rips}}(Q)\big) \le d_H(P, Q).$$

The message is: if we perturb the point cloud slightly (for instance due to sampling noise), the persistence diagrams change only slightly.